RICE UNIVERSITY

The Influence of the Tone of Feedback Prompts on the Learning Behavior and Satisfaction of University Students in a Multiple Cue Probability Learning Task

by

Sebastian Thomas

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

Master of Arts

APPROVED, THESIS COMMITTEE:

David Lane, Associate Professor,
Psychology, Statistics, and
Management, Chair

Michael Byrne, Associate Professor,
Psychology

Phillip Kortum, Professor-in-the-
Practice, Psychology

Houston, Texas
October 2010

ABSTRACT

The Influence of the Tone of Feedback Prompts on the Learning Behavior and Satisfaction of University Students in a Multiple Cue Probability Learning Task

by

Sebastian Thomas

Previous research has shown that feedback tone affects users' perceptions of computer systems. This study tested the generality of this finding and explored possible interactions of feedback tone with feedback validity and user gender. The task was a multiple cue probability learning (MCPL) problem. Experiment 1was used to establish an appropriate level of task difficulty and ensure the effectiveness of cognitive feedback. In Experiment 2, cognitive feedback validity and feedback tone were manipulated as within-subjects variables. Women improved substantially over blocks of trials in both tone conditions whereas men improved only in the polite condition. Most women preferred polite feedback whereas most men preferred the opposite. These results extend the range of tasks in which feedback tone has been shown to affect users' reactions to interfaces. These results suggest dissociation between performance and preference as men improved more with polite feedback although they preferred direct feedback.

Acknowledgements

I would firstly like to thank my advisor David Lane, whose guidance and feedback during all phases of this research was instrumental in its completion. I would also like to thank my committee members Michael Byrne and Phillip Kortum, both of whom provided insight that enhanced not only the quality of this research, but also provided direction for future research projects. Finally I would like to thank my research participants who were kind enough to participate in this project.

Table of Contents

Introduction 1
Experiment 1 9
Method 9
Subjects 9
Materials 9
Design 11
Procedure 12
Results and Discussion 14
Accuracy 15
Achievement 16
Subjective Measures 17
Experiment 2 18
Method 19
Subjects 19
Materials 19
Design 21
Procedure 21
Results and Discussion 21
Preferred Type of Feedback 22
Learning 22
Satisfaction 27
References 31

List of Tables

Table 1: Feedback prompts in experiments 1 and 2 11

Table 2: Examples of direct and polite feedback prompts used in experiment 2 21

Table 3: Means of linear components of feedback tone by order and gender. Negative values indicate a decrease in errors over blocks. 24

The Influence of the Tone of Feedback Prompts on the Learning Behavior and Satisfaction of University Students in a Multiple Cue Probability Learning Task

Formal teaching in western society can largely be characterized by instructional methodologies in which the instructor seeks to transfer knowledge to the learner via instruction, lectures etc. Despite this tradition, an increasingly dominant view of how students learn has centered on active learning. One of the earliest proponents of this approach was Piaget, whose theory of cognitive development held that learning is best served by understanding how a learner constructs knowledge rather than through repetition and copying (Piaget, 1983). The question of how constructivist methods affect learning and motivation has been a fairly active area of research and the use of learning technologies and simulations have played a large role in these research undertakings. This enthusiasm for constructivist methodology notwithstanding, researchers have to date had mixed results when it comes to demonstrating the superiority of this approach (see Mayer 2004 for a review of studies in this field).

Mayer (2004) asserts that a part of the problem with many current applications of constructivist educational approaches is the misguided view that active learning is attained through "active methods of instruction." In its purest form, this view, requires learners to discover knowledge on their own, through exploration and with minimal or no input from an instructor. Contrary to this view, Mayer (2004) argues that research should focus on guided discovery learning where active learning is facilitated with more involvement by the instructor. In support of this view, Kirschner, Sweller and Clark (2006) in their review of studies that tested the minimal guidance hypothesis conclude

that there is no evidence to support this approach. One demonstration of this is a study done by Klahr and Nigam (2004) with fourth and fifth grade children where it was found that performance on a control-of-variable task (a task designed to teach the logic behind creating valid experiments) was superior for children in the instruction condition compared to those in the pure discovery condition. These researchers also found that the transfer to a different task was equivalent for both groups. In support of Mayer's argument these researchers also point out that children in both groups engaged in active learning with the only difference being that this was supplemented by instruction in the instruction condition.

If one accepts the benefits of guidance and structure, then it is natural to ask what form of guidance is most useful to the learner. There is, for instance, some evidence that the phrasing of feedback influences user behavior in learning simulations. For example, Swaak *et al.* (2004) found that subjects did not learn very much from a computer simulation and surmised that a part of the reason for this failure was the directive language of the instructions given to subjects. Mayer, Fennell, Farmer and Campbell (2004) found that students had better learning outcomes when narration in a science application was personalized and conversational rather than formal.

The current research explored the effect of the tone of the feedback given to learners on a computer-based learning task. Of particular interest is a comparison of a polite tone with a more direct tone. What follows is a brief discussion of a theory of politeness and how it has been used in learning applications.

Politeness represents a significant aspect of how human beings interact with each other. Brown and Levinson (1987) developed a cross-cultural theory that sought to

describe how politeness is used in social interactions. They posited that individuals try to maintain "face" when they interact with each other. "Face" is from the folk term "losing face" and the theory holds that individuals try to manage two types of faces in social interactions: positive face and negative face. Positive face refers to the desire of individuals to be desirable to those they view as being important while negative face refers to the desire for individuals to not be impeded by others.

Within the MCPL paradigm used here, a feedback prompt such as "You are over weighting cue 1, you must focus more attention on the other cues" would be viewed as impolite because it entails two different face threats. First, there is a threat to positive face in the first part of the prompt which criticizes the user's performance and second, there is a threat to negative face in the second half of the prompt which prescribes a specific action that the user "must" take to rectify their mistake. The greater the number of face threats in a statement are the more impolite it is viewed as being.

Brown and Levinson's theory details a variety of strategies with regards to how politeness is used in social interactions, three of which are relevant to this research. The first strategy is known as "bald-on-record" and is one where no attempt is made to minimize face threatening actions relying instead on direct expressions. An example of this strategy in use would be the statement "You are highly under weighting the impact of Test 1, you must focus more on this test." The second strategy is conventional indirectness an example of which would be the statement "The system indicates that the entered scores highly overweight the impact of test 1." By referring to the "system" the speaker indirectly sends the message to the person being addressed to adjust their behavior. The message sent to the addressee goes beyond the literal meaning of the

sentence, but is understood through convention. The third strategy is known as joint goal and is one in which the speaker employs positive politeness by phrasing statements as joint goals. An example of this would be the statement: "It looks like we are slightly over weighting Test 1; we should pay more attention to the other tests." Here the speaker minimizes face threat by sharing in the responsibility of the addressee's actions and suggests, rather than prescribes corrective action. There are a variety of factors that account for the extent to which individuals use the various strategies when interacting with each other. The theory makes particular use of power and the closeness of the relationship between the individuals. In a tutor-student relationship the power typically resides with the tutor and this will affect how politeness strategies are utilized.

Politeness theory has implications for the design of teaching applications. Research in affective computing has shown that users' impressions of the computers they interact with can be changed by the type of language used in the prompts given by the computer. Nass *et al.* (1995) compared a computer system that used dominant-language prompts to one that used submissive-language prompts. These researchers found that subjects were more likely to prefer the computer whose personality was closer to their own personality. In this study, Nass *et al.* (1995) categorized personality as being either dominant or submissive based on a subscale from the Bem sex-role inventory (Bem, 1974). This scale has been found to correlate with other measures of dominance. Subjects scoring higher on dominance preferred direct feedback whereas those scoring lower preferred submissive prompts.

Mayer *et al.* (2006) used this theory to develop polite statements that could be used for computer-based tutors to increase the social sensitivity of educational software.

They found that subjects were able to discriminate between different levels of politeness among a set of computer application style prompts. Moore *et al.* (2004) applied principles from Brown and Levinson's theory to develop a model that could be used to generate tutorial feedback for a basic electronics tutorial that was comparable to the feedback provided by human tutors. As further support of the importance of social cues when providing feedback, Klein, Moon and Picard (2002) found that subjects used a frustrating computer longer when they were able to interact with an electronic agent that provided "active emotion support" via onscreen text.

The relationship between learning outcomes and feedback tone has not been studied extensively. In the one study I could find addressing this relationship, Wang *et al.* (2008) found better learning when subjects received feedback that was polite rather than direct. These researchers used a Wizard-of-Oz paradigm in which feedback in one condition was polite and sought to reduce face threat whereas in the other condition feedback was direct with minimal politeness. Subjects interacted with an online factory modeling and simulation application and were required to use it to forecast demand, develop a production plan and a process schedule for a virtual factory. In addition to better overall performance for subjects that received polite feedback, this study also found that subjects who reported that they preferred indirect help performed better in the polite feedback condition than the direct condition. This difference was not observed among subjects that reported a preference for direct feedback.

One key consideration in the research reported here was the choice of a learning task. Two criteria were essential in selecting a task: First, the task had to provide multiple opportunities for feedback, second, the task had to have an objective way of manipulating

the validity of the feedback, and third, there had to be an objective way of measuring user performance. The multiple-cue probability learning (MCPL) task was chosen because, as discussed later, it adequately satisfies these criteria.

The MCPL paradigm is best known for its relationship to the Brunswik lens model which holds that individuals adapt to their environment by learning how to predict future events from proximal cues (Brunswick, 1955). Brunswick developed this model around the philosophy of probabilistic functionalism which, in addition to recognizing this ability to predict future outcomes based on current cues, also held that these cues typically predict the outcome with less than 100% accuracy. The model is essentially a multiple regression where the values of a series of inputs can be used to predict the value of an outcome variable.

The lens model has been studied in a wide range of circumstances and in a variety of configurations. The model has been tested with children and young adults (Deffenbacher & Hamm, 1972; Lafon, Chasseigne & Mullet, 2004), and older adults (Chasseigne, Mullet & Stewart, 1997; Chasseigne *et al.*, 1999). It has been studied with a varying number of cues and different types of relationships between the criterion and the cues. Thus it is fair to say that the lens model is reasonably well understood with regards to MCPL which makes the difficulty level of the task easier to manipulate thus allowing for a greater degree of experimental control when assessing the impact of various types of feedback.

There are several advantages to using multiple cue probability learning for assessing the value of feedback. Experiments involving this model consist of a series of trials divided into multiple blocks. The achievement of judges can be measured within

blocks or across them. It is also possible to monitor the judge's performance by comparing the weights of the cues based on the judge's responses to the weights with the actual criterion values. This allows feedback to be tailored to the cues that the judge is having the most difficulty weighting correctly. Studies that utilize MCPL tasks have used a variety of approaches to providing feedback. This feedback can be categorized into three groups: outcome feedback, task information feedback, and cognitive/process feedback (Karelaia & Hogarth, 2008). Outcome feedback was characterized by Todd and Hammond (1965) as "knowledge of results" and refers to a condition where a judge is shown the correct criterion value after each trial. The distinction between task information feedback and cognitive feedback is somewhat blurred in the literature. Some researchers define task information feedback as information that is provided regarding the relationship between individual cues and the environmental values of the criterion while cognitive feedback is seen as information regarding the relationship between the individual cues and the criterion values provide by the judge (Karelaia & Hogarth, 2008). Others do not make this distinction and instead combine the two by viewing cognitive feedback as information about the relationships both environmental and for the judge (see Balzer, Doherty, & O'Connor, 1989; Todd & Hammond, 1965).

As will become clear, the only distinction that is important for this research is the one between outcome feedback and the other two, and for this reason task and cognitive feedback will both be referred to as cognitive feedback. Cognitive feedback is generally viewed as better for learning than is outcome feedback. Karelaia and Hogarth (2008) in a meta-analysis of lens model studies found that judges benefit more from information about the task rather than information about each trial. Todd and Hammond (1965) point

out that while outcome feedback may be appropriate for simple learning tasks it becomes less helpful when it requires subjects to have to associate individual responses with individual cue configurations over a large number of trials.

The literature raises several questions that this research sought to examine using manipulations of the MCPL task:

How does the tone of task specific feedback affect the learning and satisfaction of subjects?

Affective computing research has demonstrated that there are situations in which users respond to computers in ways similar to how they would respond to a human. The question addressed here is whether computer generated feedback that is polite leads to different learning outcomes and/or user satisfaction that does feedback that is less friendly and more direct.

Does the effect of feedback tone depend on the usefulness of the feedback?

Are subjects more trusting of feedback that is delivered in a direct or polite tone? Given the finding that a user is more tolerant of a frustrating computer if it provides feedback that is meant to empathize with that user (Moon & Picard, 2002), it would not be surprising if subjects are also more trusting of polite feedback. This might lead to higher assessments of the usefulness of invalid feedback when presented in a polite tone.

Are there gender differences in the effects of feedback tone?

As previously discussed, research has found an impact of polite feedback among subjects who scored differentially on the dominance sub scale of the Bem sex-role inventory (Nass *et al.*, 1995). While this research does not measure dominance directly,

there are large gender differences on the scale used by Nass *et al.* (1995). Therefore there may be gender difference in how individuals respond to polite versus direct feedback.

Experiment 1

Experiment 1 was a pilot study designed to refine the methodology and to calibrate the MCPL task. Only one variable, feedback validity, was manipulated. The purpose of this experiment was to develop a task for which subjects learn over blocks of trials and for which valid feedback was more useful than random feedback. Given these objectives, Experiment 1 was iterative in nature as the task was adjusted at various points during data collection based on the evaluation of the performance of subjects.

Method

Subjects. Subjects were recruited from the Rice undergraduate student population. These students received credit towards a course requirement for their participation. A total of 26 subjects were recruited. However, given the iterative nature of Experiment 1, this paper presents data from only the last five subjects whose data were collected after the last adjustment to the MCPL task. Subjects consisted of three females and two males and had an age range of 19 – 21 years.

Materials. The MCPL task was programmed using HTML, JavaScript and PHP. Figures 1 and 2 are screenshots of the two versions of the interface used in experiment 1. The application was run in Firefox 2 on a Macintosh computer. Two interfaces were developed for this experiment (see figure 1 and 2 for screenshots). The Interfaces differed

visually, but provided identical interaction schemes with the MCPL task, in that subjects could advance a trial by typing their prediction and then pressing the "Enter" key.

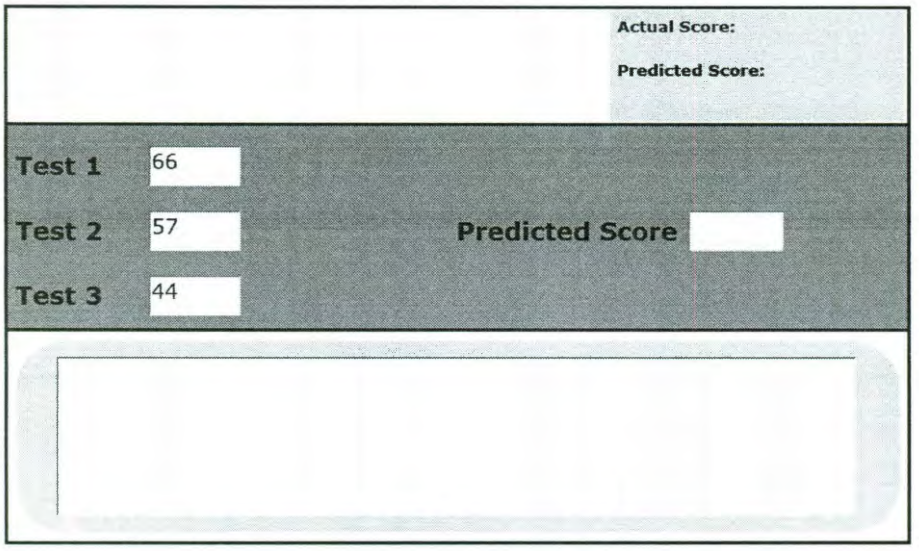

Figure 1: Interface 1

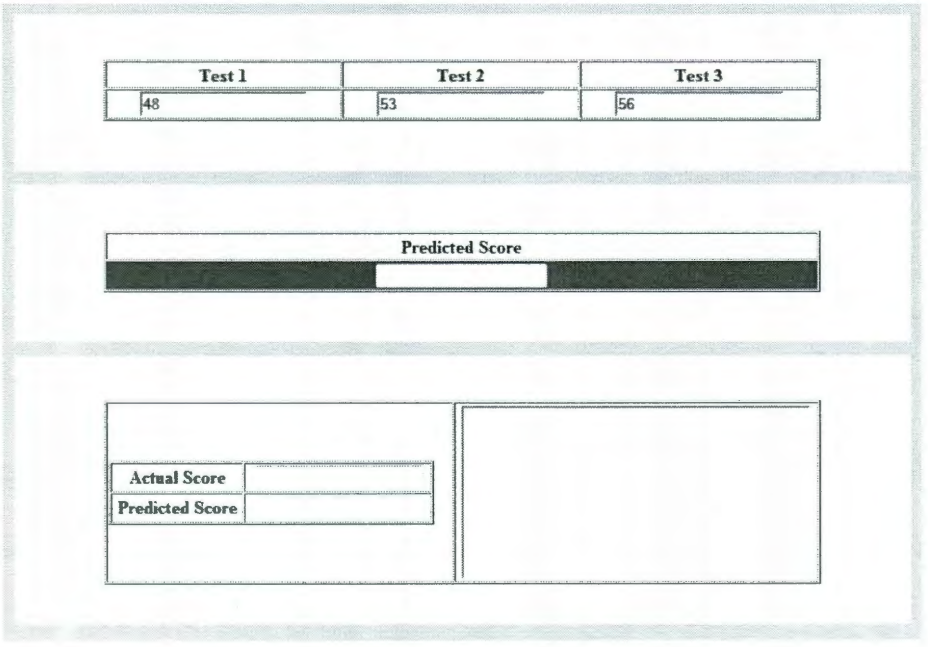

Figure 2: Interface 2

The MCPL stimuli were generated using the R statistical package and were structured such that there were three orthogonal cues with cue validities of 1, 0.5

and 0 (raw weights). These validities varied slightly for the actual experiment as all values were rounded to integers. This was done to reduce the task load on subjects by not requiring them to have to estimate fractional values. To eliminate the use of negative numbers, the means for all cues as well as the criterion were set to 50. The environmental predictability was high and ranged from .97 - .98.

Table 1 shows examples of the feedback prompts that were used in the MCPL task. Experiment 1 utilized conventional indirectness which can be considered as a politeness strategy that is somewhere in between the directness of the bald on record strategy and the positive politeness of the joint goal strategy. The System Usability Scale (SUS) was used to assess users' subjective satisfaction with the interfaces that they used.

Table 1: Feedback prompts in experiments 1 and 2

Prompt Type	Sample Statement
Bald on Record	You are under weighting the impact of test 1 you must focus more on this test.
Conventional Indirectness	The system indicates that the entered scores slightly over weight the impact of test 1, more emphasis should be place on the other tests.
Joint Goal	It looks like we are highly under weighting the impact of test 3; we should pay more attention to this test.

Design. Experiment 1 utilized an interface (2) × validity (2) design. Interface was manipulated as a between-subjects variable and each category represents one of the two interfaces in figures 1 and 2. Subjects were told that they were evaluating two interfaces so as to disguise the validity manipulation. Validity was a within-subjects variable and

contained a valid and a random feedback condition. Cognitive feedback in the valid condition was based on user inputs and thus provided accurate feedback regarding how the subject was performing on the task. In the random condition, cognitive feedback was randomized and not related to user input.

Procedure. This pilot study was used to calibrate the MCPL task. The task was framed within the context of predicting high school students' second year performance based on their score on three first year tests. Subjects were given instructions that stated that their task was to discover how important each of the first year tests was in making a prediction. It was stressed that this was a training application and that the purpose of the exercise was to evaluate two differing interfaces. It was important to not have subjects discover that the focus of the experiment was on the helpfulness of the feedback as this may have biased their learning outcomes and subjective ratings.

Subjects completed 100 trials in both a valid and a random feedback condition with a different interface used in each. Both the interfaces and conditions were counterbalanced across subjects. At the beginning of every trial, a subject was shown three first year test scores and asked to predict a second year score. After submitting a response, subject were shown what the actual score was (outcome feedback). The real b weights of the cues remained constant for all subjects in both feedback validity conditions so the three cue weights were always 1, 0.5 and 0. The order of the cues was randomized across subjects and conditions, in other words for any given set of 100 trials Test 1 may have had any of the three cue validities.

After every ten trials the application presented a summative feedback prompt regarding the subject's performance over the preceding block of trials (cognitive feedback). These prompts were either valid or random based on the condition that the subject was currently in. Feedback statements in the valid condition were based on a comparison of the real weights compared to those calculated from the subject's predictions. The displayed statement provided cognitive feedback about which cue weight (calculated based on the subject's scores) was the most different from its real cue weight. In addition to stating which cue estimate was the worst, the feedback gave subjects the real cue validity as well as the cue validity generated from the criterion values they entered. The random condition displayed a feedback statement that was identical to the one in the valid condition with the difference being that it was not based on the predictions made by the subject but rather arbitrarily selected from the set of potential statements.

Trials were self-paced, but subjects were advised to not spend an excessive amount of time (more than four or five seconds) thinking about each of their response and to rely on their first impression. Each interface had a visual indicator that indicated when five seconds had elapsed. Subjects were instructed to use this indicator as a guide and not as a time limit. Following each session of 100 trials, subjects rated the interface they used with the SUS and indicated which tests they believed to be most and least heavily weighted. After completing the task with both interfaces and conditions, subjects were asked to indicate which interface they found to be more helpful and which they found to be more aesthetically pleasing.

Results and Discussion

As stated previously, this experiment was an exploratory one whose primary purpose was to refine the MCPL task. In earlier versions of the MCPL task there were no apparent learning/validity effects and as such the task was adjusted accordingly. Given this approach, the results presented here were from the five subjects collected after the final adjustment to the task.

The changes made to the methodology included the addition of a brief description and example about the logic of multiple regression. The term regression was never used with subjects in the experiment and the example was framed in terms of the actual task subjects were about to participate in. The cognitive feedback statement was also altered to include the actual cue validity as well as the cue validity that could be derived from the criterion values entered by the subject. An example of a feedback statement would be

> *"The system indicates that the entered scores highly over weight the impact of Test 1, more emphasis should be placed on the other tests.*
> *The entered scores predict that a 1 point change in Test 1 leads to a 0.61 point change in the second year test.*
> *A 1 point change in Test 1 actually leads to a 0 point change in the second year test."*

A total of five responses across all subjects were removed from the dataset. All five were more than three standard deviations above the 75th percentiles of the responses in subject's block of trials. While there were other values that matched this criterion, the removed values which ranged from -45 to 5,504, were deemed to be erroneous entries, a decision which is justified when this range is compared to the range of the remaining values (-28 to 26). The five values were replaced with the series mean of the block in which they occurred.

Accuracy

The accuracy scores were calculated as the absolute difference between a subject's predictions and the actual criterion values. As can be seen in Figure 3, performance in the valid feedback condition was superior to that of the random condition. Although subjects were successful in reducing the differences between their predictions and the actual criterion values for both conditions, this occurred to a greater degree in the valid condition. There was fairly strong evidence for learning in the valid condition as tested by the linear component of trend, $t(4) = 2.72$, $p = .03$, one-tailed, $d = 1.22$. There was less evidence for learning in the random condition $t(4) = 1.39$, $p = .12$ one-tailed, $d = 0.62$. Although the interaction did not approach significance, $t(4) = .63$, $p = .28$, one-tailed, $d = 0.28$ However, every subject did better on the last block of trials in the valid condition than they did in the random condition. The difference between means on the last block was significant, $t(4) = 4.85$, $p = .004$, one-tailed, $d = 2.17$.

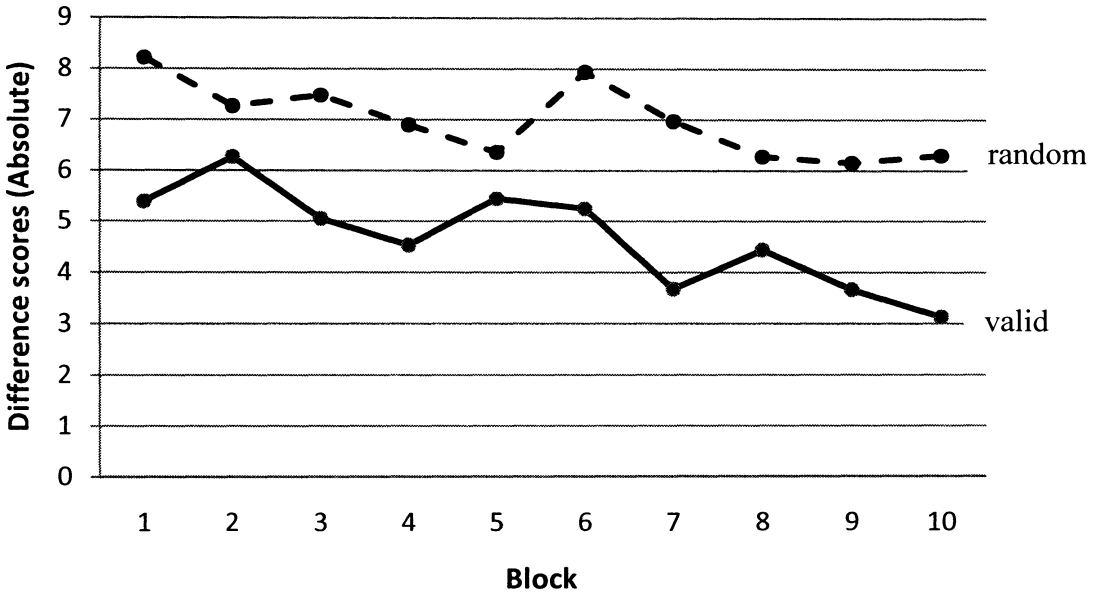

Figure 3 Absolute mean difference scores in the valid and random conditions

Achievement

The correlation between subjects' predictions and the actual criterion scores were used to determine the differences in achievement between the valid and random conditions. Figure 4 illustrates the trends for these correlations in both the valid and random conditions. Subjects generally had higher achievement in the valid condition (r = .70 to .93) than in the random condition (r = .38 to .70). Achievement improved in both conditions over the course of the blocks of trials, however as was the case with accuracy, there was stronger evidence for this improvement in the valid condition. The linear component within the valid condition was significant with a large effect size, $t(4) = 3.34$, one-tailed, $p = .02$, $d = 1.49$. This compares with the linear component in the random condition, one-tailed, $t(4) = 1.79$, $p = .075$, one-tailed, $d = .80$. While the difference between conditions was not significant $t(4) = .16$, one-tailed, $p = .44$, $d = .07$, all subjects

had a higher level of achievement in the last block of valid trials compared to the random condition $t(4) = 3.86$, $p = .01$, one tailed, $d = 1.73$.

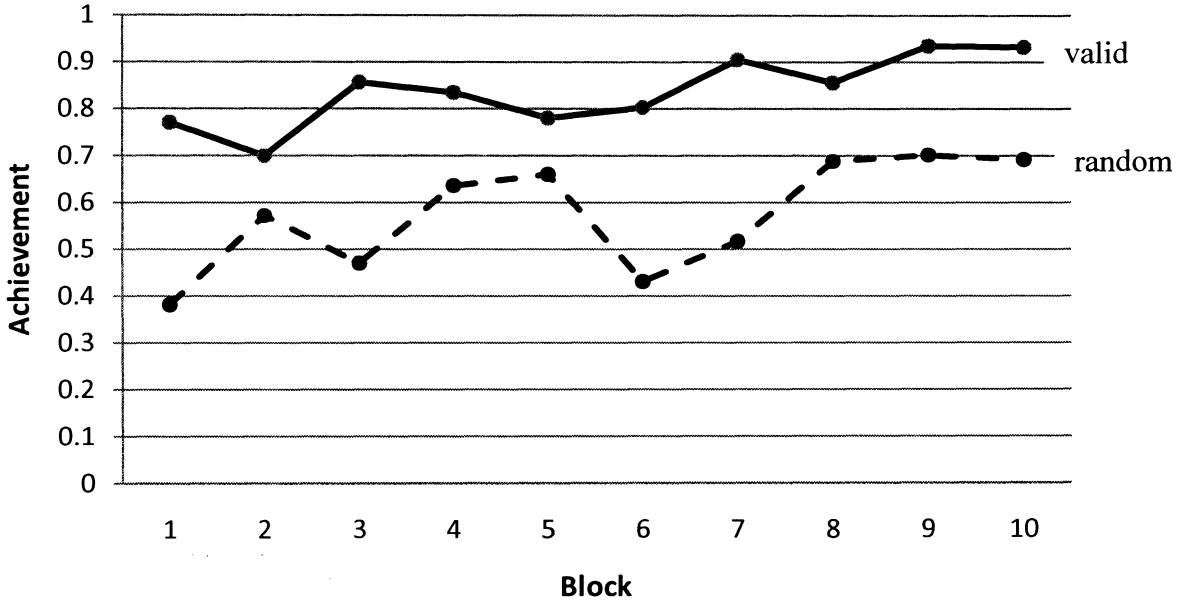

Figure 4 Achievement of subjects over ten blocks of trials for valid and random condition

Subjective Measures

When asked which interface they found to be most valid, all five subjects selected the interface from the valid condition. As illustrated in Table 1, there was very little variability in the SUS scores of both interfaces across both feedback conditions. These scores suggest that the task is difficult regardless of whether valid feedback is given or not. Although the interfaces were visually different (four of the five subjects stated that they preferred the look of interface 1), the interactions were identical for both.

Table 2 Mean SUS scores for both interfaces in the valid and random condition.

	SUS mean score
Valid	
Interface 1	48.33
Interface 2	43.75
Random	
Interface 1	46.25
Interface 2	47.5

On the basis of these results, I felt confident enough that subjects are able to use valid cognitive feedback to improve their performance in an MCPL task that I decided to begin the next experiment. While the achievement coefficients and difference scores indicated that subjects were able to improve their performance in both conditions, this improvement was generally better in the valid condition as evidenced by the effect sizes for the linear contrasts in both conditions. It should be noted that subjects are provided with accurate outcome feedback in both conditions which could account for the improvement in the random condition. In terms of subjective satisfaction, SUS scores were fairly low across all conditions, suggesting that even if subjects improved their performance in the MCPL task this did not positively impact their view of the either interface.

Experiment 2

The aim of Experiment 2 was to examine the effect of politeness in the cognitive feedback prompts on learning and satisfaction and how any potential effects might vary with gender. This experiment introduced a further manipulation of the cognitive feedback provided in the MCPL task where statements were phrased to have a direct or polite tone.

The direct feedback prompts were phrased as bald-on-record statements while the polite prompts were phrased as joint-goal statements. Based on work by Mayer (2006) these statements were found to be on opposite ends of the politeness continuum. It may be instructive to note the grammatical difference between bald on record and joint goal statements. While the former is phrased in the second-person the latter is phrased in the first-person plural. The aim of this experiment was to explore the primary research questions of whether there is an impact of feedback tone on subject performance and satisfaction as well as how tone may interact with feedback validity as well as gender.

Method

Subjects. Fifty Rice undergraduate students participated in this study (29 females and 21 males) whose ages ranged from 17 – 21 years. All subjects received credit towards a course requirement for their participation.

Materials. Two additional interfaces were added for this experiment screenshots of which can be seen in Figures 5 and 6 (see design section for description of variables). Based on the findings in Experiment 1 the cue validities for the three cues were changed to 2, 1 and 0 (raw weights). This was done to make the task less difficult. Environmental predictability ranged from .97 to .98. In a further effort to simplify the task, the mean for each cue was set to zero.

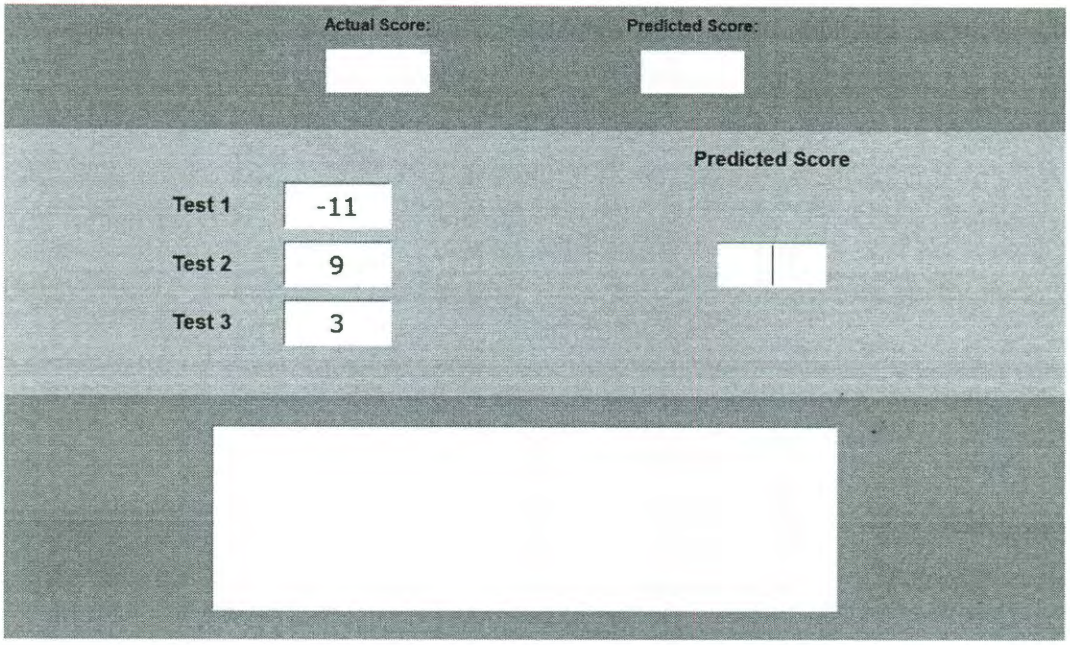

Figure 5 Interface 3

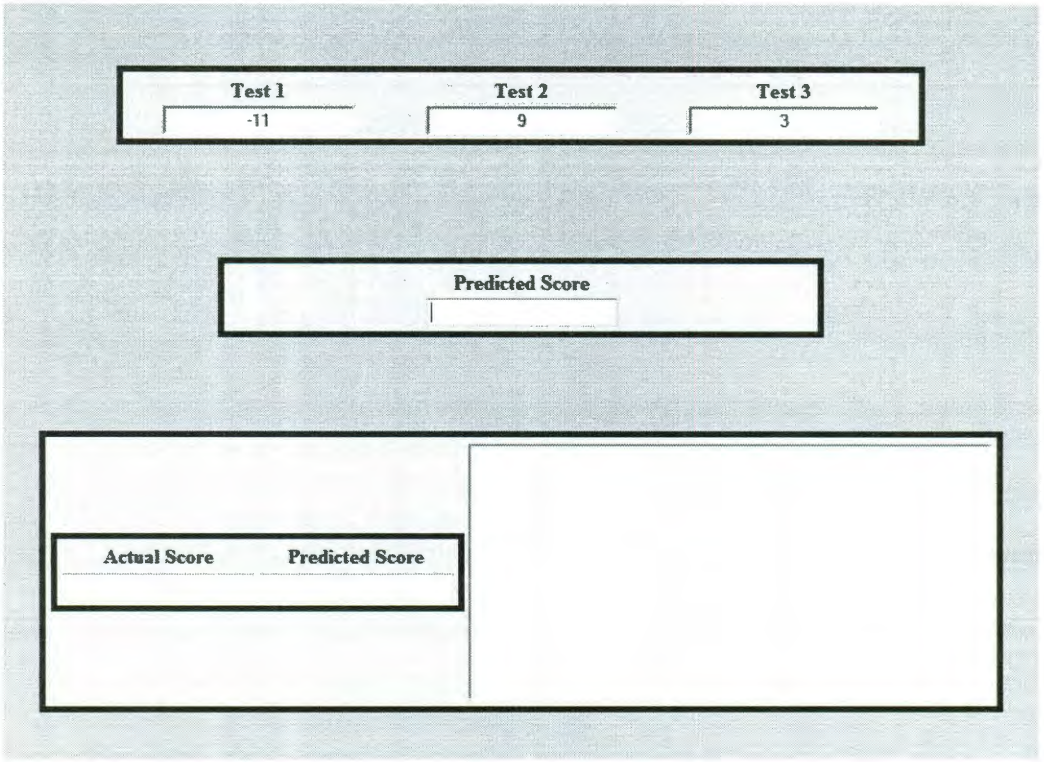

Figure 6 Interface 4

Design. Experiment 2 used an Interface (4) × Validity (2) × Feedback tone (2) within-subjects design. The additional interfaces were added to account for the feedback tone variable that was included in this experiment. In the direct feedback tone condition, cognitive feedback was phrased as bald-on-record style prompts while the polite condition utilized joint goal prompts.

Prompt Type	Sample Statement
Direct (bald-on-record)	You are under weighting the impact of test 1 you must focus more on this test.
Polite (joint goal)	It looks like we are highly under weighting the impact of test 3; we should pay more attention to this test.

Table 2 Examples of direct and polite feedback prompts used in experiment 2

Procedure. The addition of feedback tone as a within-subjects factor meant that subjects completed the MCPL task in four combinations of validity and tone. The order of these conditions for each subject was randomized. As in Experiment 1, subjects were told that they would be evaluating different interfaces rather than different types of feedback. Apart from these changes the procedures in the second experiment mirrored those of the first.

Results and Discussion

The primary research questions were (1) which type of feedback is preferred, (2) which type of feedback results in better learning, and (3), which type of feedback leads to better user satisfaction. Possible gender differences in these effects were also of considerable

interest.

Preferred Type of Feedback

Subjects were approximately evenly split with regard to whether they preferred the polite or the direct feedback: 47% (22/47) preferred the polite feedback whereas 53% (25/47) preferred the direct feedback. There was a gender difference such that 61% (17/28) of women preferred the polite feedback compared to only 26% (5/19) of the men. This difference was significant χ^2 (1, N = 47) = 5.38, p = .020. Three subjects did not respond to this question. Several subjects reported that they found the use of the "we" pronoun in the joint-goal style feedback prompts to be somewhat peculiar. There is little research that specifically addresses the grammatical person in the way computer feedback prompts are phrased and it is possible that this may have had an impact on how subjects rated feedback prompts. The awkwardness reported by subjects regarding the use of the first-person plural in the polite feedback prompt is in keeping with the finding from Nass *et al.* (2000) who reported that users do not typically hold anthropomorphic views of computer systems, despite behaving as though they do.

Learning

Performance was measured by calculating the absolute difference between subjects' judgments and those predicted by the optimal regression model. The median of the 10 data points within each block of trials for each subject was used as the score for that subject for that block. The median is a robust measure that is not greatly influenced by individual values and was thus deemed a better choice than the mean.

The first block of trials was treated as practice and was therefore excluded from all analyses. There were fifteen responses that were not included in the analysis. Eight of these were blank cells and the remaining seven were entries that exceeded the acceptable range for valid responses (-100 to 100).

Figure 7 illustrates that subjects' predictions became more accurate over blocks of trials for both the valid and random feedback conditions and that there was greater improvement for the valid than for the random feedback trials. The Validity x Block (linear) interaction was significant, $F(1, 48) = 4.97$, $p = .030$ indicating that the valid trials led to better learning than did the invalid trials. The simple effect for valid feedback (linear component) was significant, $t(49) = 6.50$, $p < .001$, $d = 0.92$, as was the simple effect for random feedback (linear component), $t(49) = 2.52$, $p = .015$, $d = 0.36$. Although it may appear counterintuitive that subjects were able to improve their performance in the random-feedback condition, this likely occurred because accurate outcome feedback was provided in both conditions and only the cognitive feedback was varied between conditions.

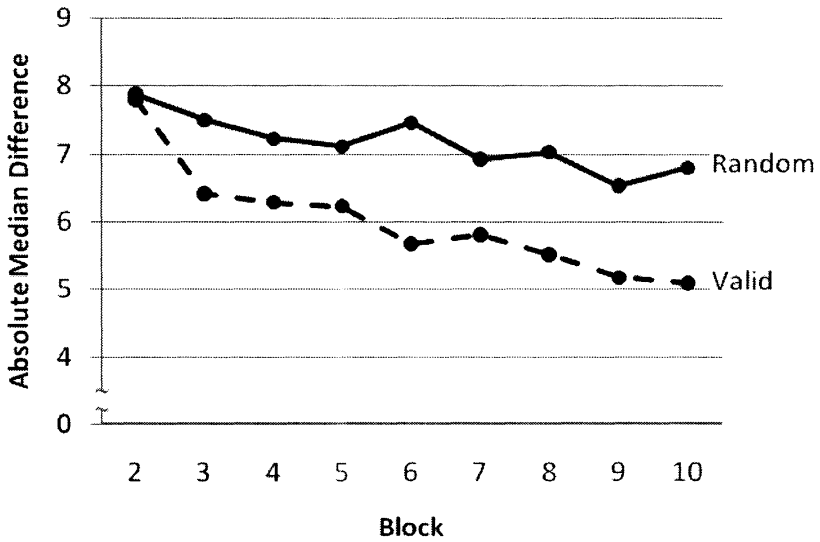

Figure 7 Absolute median difference of performance across blocks by feedback validity

As can be seen in Figure 8, women improved substantially over blocks of trials in both the polite and the direct conditions whereas men improved over blocks only in the polite condition. The Gender x Politeness x Block (linear) interaction $F(1, 48) = 6.49, p = .014$, was significant. The Politeness x Block (linear) interaction was significant for men, $F(1, 20) = 20.45, p < .001$, but not for women, $F(1, 28) = 0.10, p = .75$. There is a caveat attached to this result in that there is the possibility of an order effect. The order in which subjects were presented with polite versus direct feedback was randomized rather than counterbalanced and it turned out that 33 subjects (14 males, 19 females) received polite feedback first, while 17 subjects (7 males, 10 females) received direct feedback in their first condition. The Gender x Politeness x Block interaction is nevertheless supported by the finding that there were no significant interactions between gender and order for either of the linear components of polite or direct feedback: Polite (linear) $F(1, 49) = 0.33, p = .57$; Direct (linear) $F(1, 49) = 0.97, p = .33$. Table 3 illustrates this point as the patterns of linear component means are similar to each other and by extension to the overall Gender x Politeness x Block interaction.

Table 3 Means of linear components of feedback tone by order and gender. Negative values indicate a decrease in errors over blocks.

		Linear Components (Mean)	
	Order	**Polite**	**Direct**
Male	*Polite first*	-17.48	2.29
	Direct first	-14.54	7.18
Female	*Polite first*	-17.57	-14.21
	Direct first	-21.42	-22.9

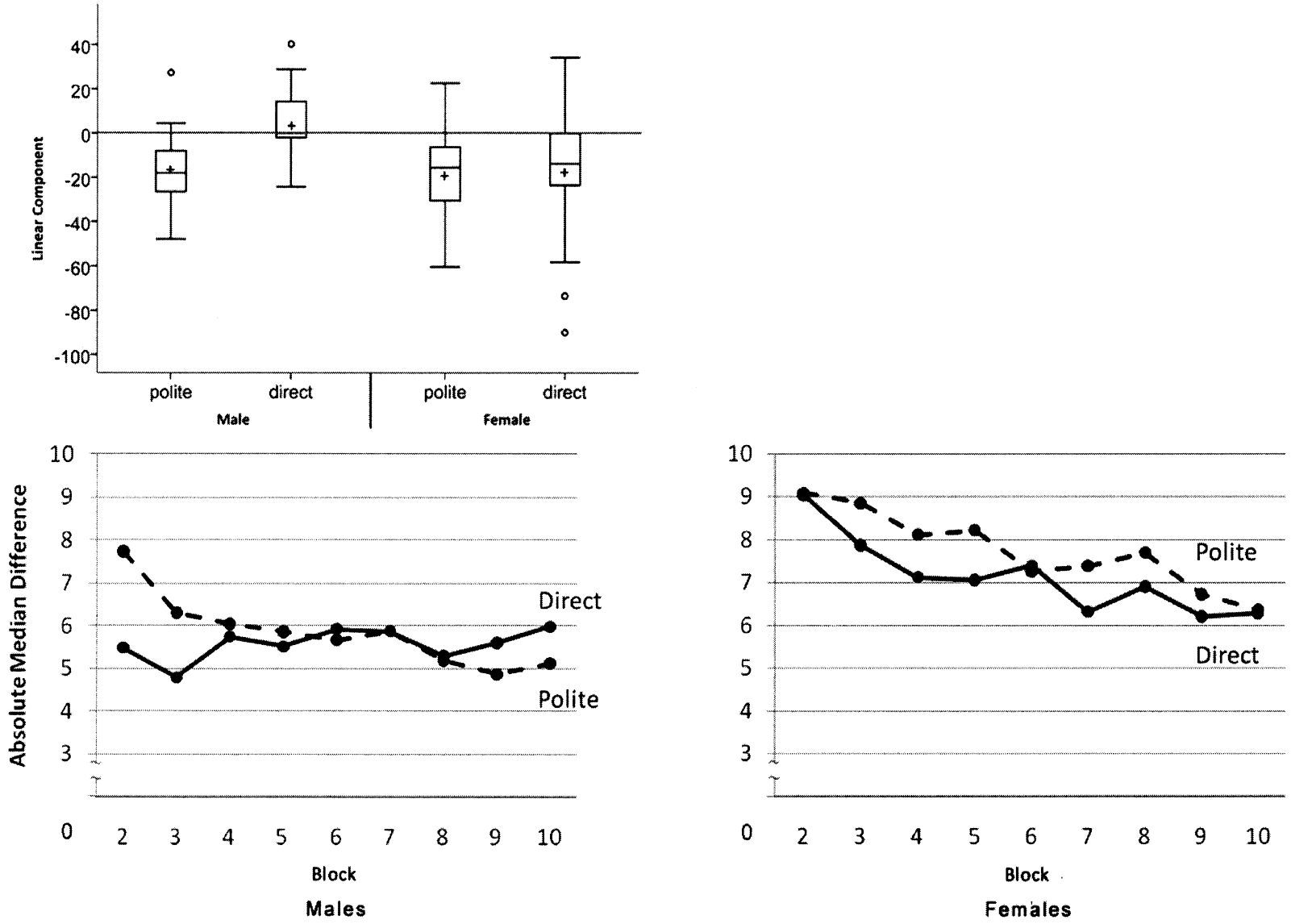

Figure 8. Upper portion: Box plots of the linear component of Blocks as a function of gender and politeness. Lower portion: Absolute median difference of performance across blocks by tone and gender.

The finding that men improved more with polite than with direct feedback whereas women improved with both direct and polite feedback is not easily explained and this finding would certainly need to be replicated before firm conclusions should be drawn. There are some studies of gender differences in how feedback is used that may be relevant. Fagot (1985) in a study of very young children (21-25 months) found that boys did not respond to positive or negative feedback given by teachers or girls, while girls responded to both types of feedback from teachers. Roberts (1991) used social role theory to assert that men may be less responsive to evaluative feedback because they tend to be socialized in a more combative peer group and so they are more likely to ignore critical feedback. Women on the other hand tend to be socialized in a more collectively-oriented peer group and are thus more receptive to feedback in general.

There was no evidence of an interaction between feedback validity and tone as neither the Tone × Validity interaction, $F(1, 48) = 0.49$, $p = .83$) nor the Block × Tone × Validity, $F(1, 48) = 0.002$, $p = .97$, interaction approached significant.

A second measure of performance assessed whether subjects were able to correctly select which of the three tests was most heavily weighted and which was the least at the end of each condition. Responses were scored as one for a correct response and zero for an incorrect response. Figure 8 shows the proportion of subjects that correctly determined the least and most important of the three cues at the end of each condition. Subjects were better at determining the least important test than the most important one $F(1,48) = 43.35$, $p < .001$. Also apparent from Figure 8 is that subjects were more accurate at selecting the least important test in the valid feedback conditions than in the random condition whereas the difference between valid and random

conditions was negligible for the choice of the most important cue. The Importance x Validity interaction was a significant, $F(1,48) = 4.81$, $p = .03$. There was no evidence of an effect of tone nor did tone significantly interact with any variable (lowest $p = .14$). There were no apparent gender difference with respect to importance or its interaction with any of the other variables (lowest p = .54).

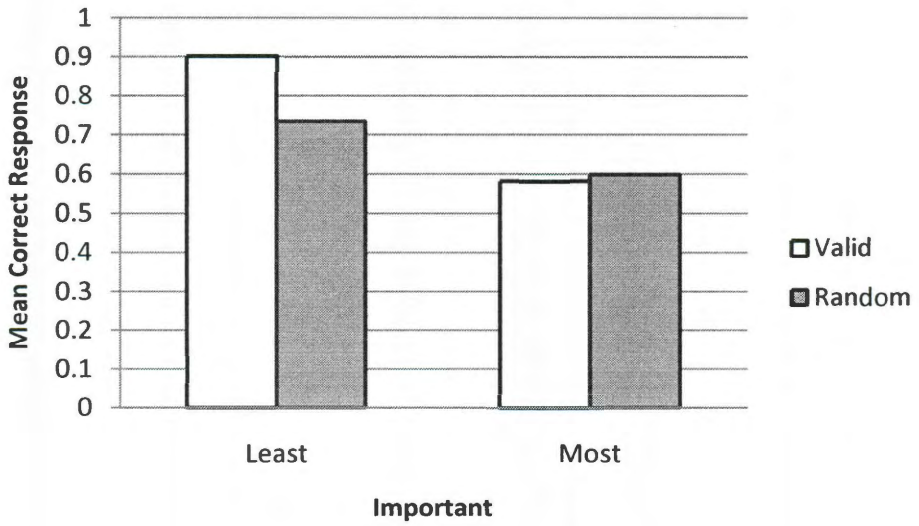

Figure 8 Proportions of correct responses by test importance and feedback validity

Satisfaction

Subjects completed the SUS scale for each of the four conditions. Figure 9 displays the means for these conditions broken down by gender. The means scores were all fairly close to each other across both genders hovering at around 60. Although the highest mean score occurred in the valid and direct condition for both genders this was not found to be statistically significant.

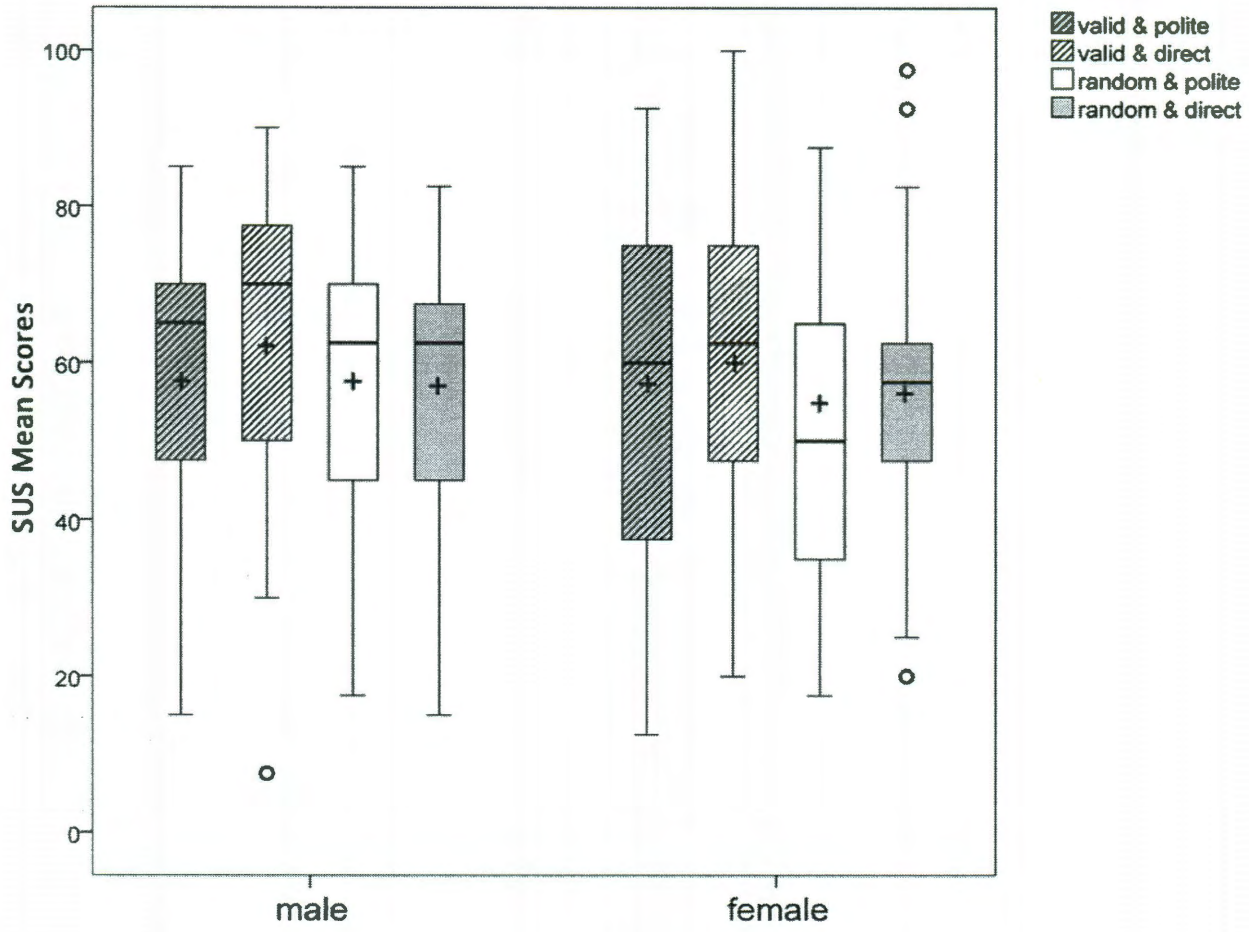

Figure 9 Mean SUS scores by feedback condition and gender

Overall, the SUS scores were higher for the direct feedback (M = 59.03) than for the polite feedback (M = 55.63). However, this difference should be interpreted cautiously since the difference did not reach conventional levels of significance, $F(1, 48) = 3.83$, $p = .056$.

None of the other effects or interactions approached significance (lowest $p = .15$).

The SUS results do not match those found for preferred feedback where there was a significant difference between women (who tended to prefer polite feedback) and men (who tended to prefer direct feedback). This difference in results is most likely attributable to differences in the measures used. The SUS scale is more of a measure of

the interface and how well it could be utilized to complete the prescribed task. Feedback prompts, in this case, are only a part of the overall system and so are not directly measured by the SUS scale. Conversely, the direct measure of feedback preference used in this study fails to address some of the nuances that may exist with the two types of prompts. There is, for instance, the previously discussed issue of the polite feedback being phrased in the first-person plural and the direct feedback being phrased in the second-person. One recommendation for future directions that arises out of this study is to further explore these differences.

The results from Experiments 1 and 2 provide support for the notion that the tone of feedback affects the way in which users perceive and perform with a learning application. There is also evidence that gender plays a role in these effects. Women improved their performance in both the direct and polite feedback conditions while men were only able to improve in the polite condition. When asked which interface they liked the most, a majority of women selected one with polite feedback while a most men selected one with direct feedback. Future research in this area should continue to experiment with different types of feedback prompts.

Since the two tones used here differed in terms of whether the feedback was first person versus second person, it remains for future research to determine whether other manipulations of tone would have the same effect and whether the critical variable is tone or grammatical form. It would also be instructive to explore more nuanced measures of subjects' feedback prompt ratings. Feedback tone should also continue to be explored across a variety of tasks and learning situations. The unexpected finding of men preferring direct feedback despite having better performance with polite feedback is one

that certainly merits further examination. Feedback is important to learning and designers of teaching applications should consider not only the content of the feedback these systems will provide, but also the tone that is used to communicate this feedback. As designers strive to improve user experience and create intelligent virtual agents whose purpose is to assist, some thought should be given to how these agents will communicate with the learner. It is apparent that the social norms that govern communication between a human tutor and a student apply to some degree even when the tutor is a virtual agent. As the tone of feedback may impact not only user's perception of the system but also learning outcomes a learning application that is designed with these norms in mind is likely to be superior to one that does not.

References

Balzer, W. K., Doherty, M. E., & O'Connor, R. Jr. (1989). Effects of cognitive feedback on performance. *Psychological Bulletin, 106*, 410 – 433.

Bem S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology, 42*, 155 – 162.

Brown, P., Levinson, S. C. (1987). Politeness: Some universals in language use. Cambridge University Press, New York.

Brunswick, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychology Review, 62*, 193 – 217.

Chasseigne, G., Grau, S., Mullet, E., & Cama, V. (1999). How well do elderly people cope with uncertainty in a learning? *Acta Psychologica, 103*, 229 – 238.

Chasseigne, G., Mullet, E., & Stewart, T. R. (1997). Aging and multiple cue probability learning: The case of inverse relationships. *Acta Psychologica, 97*, 235 – 252.

Cuqlock-Knopp, V. G., Wilkins, C. A., & Torgerson, W. S. (1991). Multiple cue probability learning and the design of information displays for multiple tasks. In D. L. Damos (Ed.) *Multiple-task Performance* (pp. 139 – 152). London: Taylor & Francis Inc.

Deffenbacher, K. A., & Hamm, N. H. (1972). An application of Brunswik's lens model to developmental changes in probability learning. *Developmental Psychology, 6,* 508 – 519.

Fagot, B. I. (1985). Beyond the reinforcement principle: Another step toward understanding sex roles. *Developmental Psychology*, *21*, 1097–1104.

Karelaia, N., & Hogarth, R. M. (2008) Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin, 134,* 404 – 426.

Kirschner, P. A., Sweller, J., Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychology, 41,* 75 – 86.

Klahr, D., & Nigam, M. (2004). The equivalence of learning paths in early science instruction: Effects of direct instruction and discovery learning. *Psychological Science, 15*, 661-667.

Klein, J., Moon, Y., & Picard, R. W. (2002). This computer responds to user frustration: Theory, design, and results. *Interacting with Computers, 14,* 119 – 140.

Lafon, P., Chasseigne, G., & Mullet, E. (2004). Functional learning among children, adolescents and young adults. *Journal of Experimental Child Psychology, 88,* 334 – 347.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist, 59,* 14 – 19.

Mayer, R. E., Fennell, S., Farmer, L., & Campbell, J. (2004). A personalization effect in multimedia learning: Students learn better when words are in conversational style rather than formal style. *Journal of Educational Psychology, 96,* 389 – 395.

Mayer, R. E., Johnson, W. L., Shaw, E., & Sandhu, S. (2006). Constructing computer-based tutors that are socially sensitive: Politeness in educational software. *International Journal of Human-Computer Studies, 64,* 36 – 42.

Moore, J. D., Porayska-Pomsta, K., Varges, S., & Zinn, C. (2004). Generating Tutorial feedback with affect. In *Proceedings of the 17th International Florida Artificial Intelligence Research Society conference.*

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56,* 81 – 103.

Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, D. C. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies, 43,* 223 – 239.

Piaget, J. (1983). Piaget's Theory. In P. H. Mussen (Series Ed.) & W. Kessen (Vol. Ed.), *Handbook of Child Psychology: Vol. 1. History, Theory, and Methods* (4th ed., pp. 103 - 128). New York: John Wiley & Sons.

Roberts, T. A. (1991) Gender and the Influence of Evaluations on Self-Assessments in Achievement Settings, *Pyschological Bulletin,* 109, 297-308.

Swaak, J., de Jong, T. & van Joolingen, W. R. (2004). The effects of discovery learning and expository instruction on the acquisition of definitional and intuitive knowledge. *Journal of Computer Assisted Learning 20*, 225–234.

Todd, F. J., & Hammond, K. R. (1965). Differential feedback in two multiple-cue probability learning tasks. *Behavioral Science, 10*, 429–435.

Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer Studies,* 66, 98-112.

White, C. M. & Koehler, D. J. (2007). Choice strategies in multiple-cue probability learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 757 – 768.

Zhang, J., Chen, Q., Sun, Y., & Reid, D. J. (2004). Triple scheme of learning support design for scientific discovery learning based on computer simulation: experimental research. *Journal of Computer Assisted Learning 20*, 269-282.